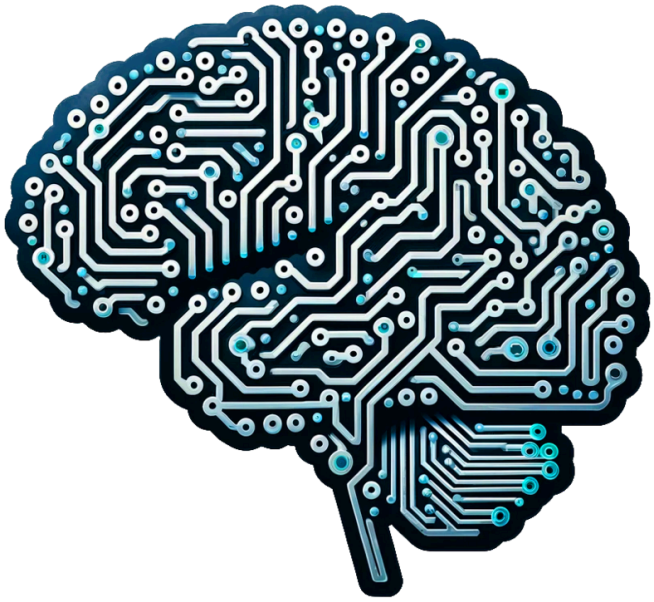
Introducción a ML y Generative AI



Taller de programación #2

Descripción

En este taller, los estudiantes aplicarán varias técnicas de preprocesamiento de datos utilizando Python en un entorno Jupyter Notebook dentro de Visual Studio Code (VSCode). Trabajarán con el Titanic dataset para limpiar, transformar, y escalar los datos en preparación para un modelo de Machine Learning.

1. Herramientas :

- **Python**: versión 3.11
- **VSCode**: con soporte para Jupyter Notebooks
- **Bibliotecas**: pandas, numpy, sklearn, matplotlib (para visualización opcional)

2. Explorando el Titanic Dataset

Usando pandas, procederemos a realizar una exploración inicial. Para ello, deberás realizar los siguientes pasos:

- Abrir VSCode y crear un notebook.
- Cargar el dataset en un DataFrame de pandas.
- Mostrar las primeras 5 filas y obtener un resumen de las características usando .info() y .describe().

3. Manejo de Datos Faltantes

Como hemos visto en nuestro análisis básico, existen valores nulos, como en la columna edad. Por esta razón, debemos identificar y manejar los datos faltantes en el dataset.

- Detectar las columnas con valores faltantes usando isnull().sum().
- Imputar los valores faltantes en la columna Age con la mediana.
- Eliminar la columna Cabin por tener demasiados valores faltantes.
- Imputar los valores faltantes de Embarked usando la moda.

4. Manejo de Datos Categóricos

La mayoría de los algoritmos de machine learning y técnicas estadísticas requieren entradas numéricas. Los datos categóricos deben transformarse en un formato numérico para que estos modelos puedan procesarlos y aprender de ellos.

- Aplicar Label Encoding a la columna Sex.
- Aplicar One-Hot Encoding a la columna Embarked.

5. Detección y Manejo de Outliers

Existen datos que se encuentran significativamente alejados del resto del conjunto (outliers). Los outliers pueden influir desproporcionadamente en la función de costo de algunos algoritmos de machine learning, como la regresión lineal, lo que puede llevar a un modelo que no representa adecuadamente la relación entre las variables para la mayoría de los datos. Vamos a averiguar si existen outliers en la columna Fare, para esto necesitamos:

- Visualizar la distribución de Fare usando un histograma o boxplot.
- Decidir si es necesario transformar o eliminar outliers.

6. Escalado de Datos Numéricos

Para culminar este taller, aplicaremos técnicas de escalado a las columnas numéricas Age y Fare.

- Aplicar Z-Score Scaling a las columnas Age y Fare.
- Aplicar Min-Max Scaling como alternativa.

7. Preguntas de reflexión:

- ¿Qué técnica de encoding fue más adecuada para los datos categóricos?
- ¿Cómo impacta el escalado en los datos?
- ¿Qué decisiones tomarías si hubiera más datos faltantes o outliers?
- ¿Qué otros tipos de encoding serían útiles para el Titanic dataset?

Marcel Mauricio Moran Calderon
marcel_moran41@hotmail.com

Ariel Ramos Vela
ariel.ramos97@gmail.com